Responsiveness of general health status in chronic low back pain: a comparison of the COOP Charts and the SF-36

Bronfort, G.; Bouter, L.M.

published in
Pain
1999

DOI (link to publisher)
10.1016/S0304-3959(99)00103-7
S0304395999001037 [pii]

document version
Publisher's PDF, also known as Version of record

Pain 83 (1999) 201–209

PAIN

www.elsevier.nl/locate/pain

Responsiveness of general health status in chronic low back pain: a comparison of the COOP Charts and the SF-36

Gert Bronfort[a,*], Lex M. Bouter[b]

[a]Department of Research, Wolfe-Harris Center for Clinical Studies, 2501 West 84th Street, Bloomington, MN, 55431, USA
[b]Institute for Research in Extramural Medicine and Department of Epidemiology and Biostatistics, Faculty of Medicine, Vrije Universiteit, van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands

Received 11 November 1997; received in revised form 5 January 1999; accepted 30 April 1999

Abstract

The objective of this study was to compare the responsiveness and assess the concurrent validity of two functional health status instruments, the Dartmouth COOP charts and the SF-36 in chronic low-back pain (CLBP) patients. The data came from 129 of 174 patients who participated in a randomized clinical trial of the therapeutic management of CLBP. Reliable and valid disease-specific outcomes, patient-rated low-back pain and disability, were used as external criteria (EC) to identify improved and non-improved patients. Unpaired t-statistics and receiver operating characteristic (ROC) curve calculations were used to quantify responsiveness. The two instruments had sufficient and very similar responsiveness using both EC. Comparisons between improved and non-improved patients for the COOP charts and SF-36, respectively, using pain as EC, yielded differences which translated into large effect sizes (0.8 and 0.7) ($P = 0.0008$ and 0.003). Using disability as EC, differences of moderate effect size were found (0.5 and 0.6) ($P = 0.02$ and 0.002). The ROC curve calculations using pain as EC resulted in areas under the curve of 0.76 (95% CI: 0.64, 0.88) for the COOP charts, and 0.74 (95% CI: 0.60, 0.88) for the SF-36. The corresponding areas using disability as EC were 0.67 (95% CI: 0.55, 0.79) and 0.72 (95% CI: 0.60, 0.84). The best cut-off point in both instruments for differentiating between improved and non-improved patients was approximately six percentage points. The constructs of functional health status, as reflected in the global scores of the two instruments, are highly correlated ($r = 0.82$). Six of the instruments' nine dimensions are moderately to highly correlated ($r = 0.52$ to 0.86), and the overall canonical correlation was high ($R = 0.9$). In conclusion, both instruments seem equally suitable for use as outcome measures in clinical trials on CLBP. The COOP charts are faster to fill out and score. © 1999 International Association for the Study of Pain. Published by Elsevier Science B.V.

Keywords: Low-back pain; Functional health status; COOP charts; SF-36; Responsiveness; Randomized clinical trials

1. Introduction

Psychosocial factors play an important role in influencing the course of chronic low-back pain (CLBP) (Deyo and Diehl, 1983). Instruments to measure functional health status, including physical, emotional and social functioning, have been around for many years, but in general have been too lengthy for both clinical and research purposes (Deyo and Diehl, 1983). Within the last 10 years briefer functional health status questionnaires have been used with increasing frequency as important outcomes in randomized clinical trials (RCTs) (Anderson et al., 1993). Two popular instruments are the SF-36 and the COOP charts. An early version of the SF-36 (the MOS short form) has been utilized in the large Medical Outcome study in over 20 000 patients and was proven to be reliable, valid and capable of creating distinct health profiles for patients with different chronic conditions, including CLBP (Stewart et al., 1989).

The cooperative (COOP) chart system was developed by the Dartmouth COOP Project Network of community practices for use in primary care settings (Nelson et al., 1990a; 1983; Shigemoto, 1990). This instrument has been shown to be reliable and valid when tested on several thousands of patients in diverse primary care settings in United States, Europe and Japan (Nelson et al., 1987, 1990; Meyboom-de Jong et al., 1990; Kinnersley et al., 1994). The COOP charts and the MOS short form have been compared and have shown similar sensitivity in detecting the effects of several prevalent diseases, such as heart disease and depression, on functional health status (Landgraf, 1990).

Besides reproducibility and validity, another very important property of an outcomes instrument is the ability to capture clinically important change in specific disorders

* Corresponding author. Tel.: +1-612-885-5413; fax: +1-612-942-8456.
E-mail address: gbronfort@aol.com (G. Bronfort)

0304-3959/99/$20.00
PII: S0304-3959(99)00103-7

(Deyo et al., 1991; van Bennekom et al., 1996). This sensitivity to change in a patient's condition, as defined by a meaningful external criterion (EC), is termed responsiveness. Both the COOP charts and the SF-36 have been used internationally as important outcomes in clinical trials for a variety of clinical conditions, but their responsiveness has not been evaluated in CLBP patients. The purpose of this study was to compare responsiveness and assess concurrent validity of both instruments in a sample of CLBP patients in the context of a randomized clinical trial (RCT).

2. Methods

The main purpose of the RCT was to study the relative efficacy of three different treatment regimens for CLBP in 174 patients aged 20–60 years. Patients were recruited through newspaper advertisement. All patients were treated in a primary contact multidisciplinary outpatient clinic. There is evidence to suggest that sufficient similarity exists between patients recruited through advertising and primary clinical settings to make generalization of findings from studies like ours possible (Deyo et al., 1990; Koes et al., 1992).

The therapeutic interventions consisted of 11 weeks of spinal manipulative therapy (SMT) combined with trunk strengthening exercises or trunk stretching exercises, or trunk strengthening exercises combined with a prescription of non-steroidal anti-inflammatory medication. The results of this RCT have been reported elsewhere (Bronfort et al., 1996). One of the secondary research questions of the RCT was to compare the responsiveness and to assess the concurrent validity of two functional health status instruments, the COOP charts and the SF-36. For valid comparison the two instruments were administered in random order at baseline and after 11 weeks of therapy. Both instruments' raw scores for each dimension were transformed into percentage scores to allow for direct comparison. The percentage scores of all nine dimensions of each instrument were summed and divided by 9 to arrive at a global score. The respective global scores were thus constructed without weighting of the individual dimensions.

2.1. Responsiveness

Responsiveness is defined as the ability of an outcomes instrument to detect clinically important changes in a specific condition. The methods used to quantify responsiveness require the use of at least one valid EC of improvement.

2.2. External criteria of improvement

For the purposes of this study the external criteria (EC) for determining improved and non-improved patients were based on two low-back specific outcome measures in our trial, patient-rated low-back pain (LBP) severity and the Roland–Morris low-back disability index (RMI). Patient-rated LBP severity was recorded on an 11 box scale for ease of administration and scoring. Each box has a number with anchors at 0 denoting no symptoms and at 10 denoting the highest severity of pain. This pain scale has been shown to have reliability and validity comparable to the 10 cm visual analog scale (Jaeschke et al., 1990). The RMI consists of 24 yes/no questions describing different types of restriction in daily activities specifically due to LBP (e.g. 'because of my back, I lie down to rest more often'). This instrument has measures of reliability, validity and responsiveness similar to the much longer, highly reliable and valid sickness impact profile (SIP) from which it was derived (Roland and Morris, 1983). Both EC have been used extensively as main outcomes in RCTs on LBP (Deyo, 1986; Deyo et al., 1994) and recently been shown to be highly responsive measures in chronic LBP patients (Beurskens et al., 1996).

The difference between improved and non-improved patients had to amount to a minimal clinically important difference (MCID). MCID was defined as the smallest difference in score of a particular outcome that patients would consider beneficial and which would form the basis for changing therapeutic management in the absence of serious side effects and prohibitive costs (Jaeschke et al., 1989). Data from most studies which have addressed this issue suggest, regardless of the clinical condition, that an MCID, which usually translates into an ES of approximately 0.5, corresponds to a change in patient rated pain of approximately 8–10 percentage points (Jaeschke et al., 1989; Goldsmith et al., 1993; Juniper et al., 1994). The effect size represents a unitless standardized difference. In this example the effect size is derived at by dividing the difference in means between two groups with the standard deviation of the difference. An effect size of 1.0 is equivalent to a difference of 1 standard deviation in the sample. To assess the relative magnitude of difference in scores, Cohen identified an effect size of 0.2 as small, 0.5 as moderate and 0.8 as large (Cohen, 1988a). For both patient-rated LBP severity and the RMI, we defined the improved group to have at least 10 percentage points improvement. Non-improved (stable, unchanged) patients were defined as those having a range of change scores from -5 to $+5$ percentage points. Consequently, the rest of the patients, those with equivocal improvement (5–10 percentage points) and those showing deterioration (beyond -5 percentage points) were not included in the responsiveness analysis. These criteria for dividing patients into improved and non-improved individuals were used in connection with both of the methods for quantifying responsiveness described below.

2.3. Methods for quantifying responsiveness

Two complementary methods for quantifying responsiveness were employed: unpaired *t*-statistics (Guyatt et al., 1989) and receiver operating characteristic (ROC) curves (Deyo et al., 1991; Hanley and McNeil, 1982).

Table 1
Demographic and baseline clinical characteristics[a]

Characteristic	
No. of subjects	129
Age (year)	42.2 (9.2)
Gender (% female)	50.8
Working at full capacity at study entry (%)	91.9
Duration of current episode of low back pain (year) (median)	2.0
Pain radiation to leg (%)	52.0
Previous hospitalization for low back pain (%)	6.4
Smoker (%)	15.9
Three or more previous episodes of low back pain (%)	48.4
Low back pain score (0–10 scale)	5.4 (1.5)
Low back disability score (Roland–Morris, 0–100 scale)	34.6 (18.3)
Global general health score (COOP charts, 0–100 scale)	64.2 (12.8)
Global general health score (SF-36, 0–100 scale)	65.2 (14.1)

[a] Values are means and standard deviations (SD) unless otherwise noted.

2.4. Unpaired t-tests

Deyo and Centor (1986) and Guyatt et al. (1989) defined a responsive instrument as one that is clearly capable of differentiating between improved and non-improved patients as determined by one or more relevant EC. The greater the difference in change scores between improved and non-improved patients and the smaller the *P*-values, the greater the responsiveness. Since *P*-values are dependent on sample sizes, we divided the differences in change scores by the pooled standard deviations of those change scores to yield effect sizes for standardized comparisons among instruments. Correction for effect size estimate bias associated with small sample sizes ($n < 50$) was accomplished using the method described by Hedges and Olkin (1985).

2.5. Receiver operating characteristic (ROC) curves

Deyo et al. (1991) have suggested that health status questionnaires can be regarded as diagnostic tests capable of distinguishing between improved and unimproved patients. Based on this concept it becomes possible to construct ROC curves which describe the instrument's ability to detect improvement or lack thereof on the basis of one or more relevant and valid EC. The ROC curves display positive (sensitivity) versus false positive (1-specificity) rates for a series of cut-off points in change scores. The area under the ROC curve can be calculated and interpreted as the probability of correctly identifying improved patients from randomly selected pairs of improved and unimproved patients. An area of 0.5 is interpreted as no discriminatory accuracy and 1.0 as complete accuracy. In addition, ROC curves can be compared statistically, (Hanley and McNeil, 1983) as well as form the basis for deciding which cut-off points are best capable of discriminating between improved and non-improved patients (Hanley and McNeil, 1982).

2.6. Concurrent validity

Concurrent validity was assessed by calculating Pearson's product-moment correlation between global scores of the COOP charts and the SF-36 at baseline and after 11 weeks of treatment ($n = 129$). Canonical correlation was used as an additional procedure for assessing concurrent validity by correlating the individual sets of similar dimensions of the two instruments. This procedure involves correlation of two derived variables, each representing a weighted combination of the two sets of nine dimensions. The canonical weights, similar to the beta weights of multiple correlation, are calculated in a way that the correlation of the two derived variables are maximized (Kachigan, 1986).

2.7. Statistical analysis

The software program True Epistat™ version 5, (Richardson, Texas) was used to calculate the ROC curves, to test for instrument differences in areas under the curve, and to compare instrument sensitivity and specificity. The *t*-tests, Mann–Whitney tests, correlation coefficients, and canonical correlations were calculated using the statistical software package Statistica for Windows ver. 5.0 by Statsoft, Inc.

3. Results

3.1. Characteristics of the subjects

Seventy-six percent (132 patients) of the initial 174 patients completed the intervention part of the study (11 weeks) and were available for all main outcome assessments. COOP and SF-36 data were available at baseline and week 11 for 129 of the 132 subjects. We investigated the means, standard deviations, and minima and maxima for all patients, including drop-outs, for baseline clinical and demographic variables and all primary outcomes at the three times of assessment. We were unable to identify any patterns that suggested anything except that the data were missing completely at random (MCAR) (Little and Rubin, 1987). MCAR is the term used to indicate that the missingness is like a random sample, hence analyzing the completers has reduced power but should not be biased. This means the drop-outs represent a random sample of patients in this trial, and it is unlikely that data from these patients would have changed the main results of the study.

The main demographic and clinical characteristics of the subjects are summarized in Table 1. Sixty-seven percent had jobs outside the home; 23% were self-employed; and 3% were unemployed. Twenty-seven percent had manual labor occupations.

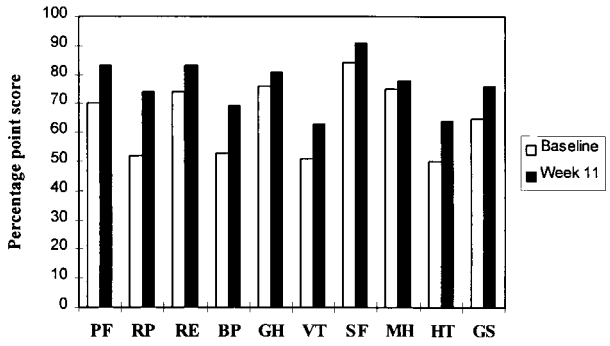

Fig. 1. SF-36 dimension score profile in percentage points at baseline and at week 11. (Higher scores are better regardless of dimension). PF, Physical functioning, RP, Role physical; RE, Role emotional; BP, Bodily pain; GH, General health; VI, Vitality; SF, Social functioning; MH, Mental health; HT, Health transition; GS, Global score.

3.2. The main therapeutic outcomes

The results of our RCT has been reported elsewhere (Bronfort et al., 1996). Clinically important and statistically significant improvement over time was evident in all the major outcome measures in the entire study population, but there were no important outcome differences between interventions. The largest improvements were seen in patient rated pain and disability, based on the mean changes from baseline to after 11 weeks of treatment, which were 22.9 (95% CI: 26.7, 19.1) and 16.4 (95% CI: 19.3, 13.4) percentage points, respectively. The improvement in global score for the SF-36 was 11.0 percentage points (95% CI: 8.8, 13.3), and for the COOP charts, 11.2 percentage points (95% CI: 9.0, 13.4). Figs. 1 and 2 display the change in COOP and SF-36 individual dimension score profiles and global scores over time. The responsiveness in terms of statistical power was greatest for patient-rated pain with the pre-post treatment difference translating into an effect size of 1.5 compared to 0.8 to 1.0 for the other three variables.

3.3. Responsiveness

Quantification of responsiveness of the two general health status instruments showed pronounced similarity with both methods of analysis using both EC. The global change score difference between improved and non-improved patients, using patient-rated pain and low-back disability as EC were all clinically important and statistically significant (Table 2). In regard to EC scores, of the 132 patients with available data, those with improvement between 5 and 10 percentage points, for LBP severity ($n = 2$ (2%)) and RMI ($n = 15$ (11%)), and those showing deterioration beyond -5 percentage points, for LBP severity ($n = 11$ (8%)) and RMI ($n = 10$ (8%)), were not included in the analysis. The ROC curve calculations are displayed in Figs. 3 and 4. Using pain as EC resulted in an area under the curve of 0.76 (95% CI: 0.64, 0.88) for the COOP charts and 0.74 (95% CI: 0.60, 0.88) for the SF-36, (difference, $P = 0.76$). The corresponding areas using low-back disability as EC were 0.67 (95% CI: 0.55, 0.79) and 0.72 (95% CI: 0.60, 0.84) (difference, $P = 0.55$). The best cut-off point for both the COOP charts and the SF-36 in differentiating between improved and non-improved patients was estimated at 6 percentage points (the data point closest to the upper left corner of the ROC graph). The sensitivity/specificity using this cut-off point were 74%/80% and 74%/65% for the COOP charts and the SF-36, respectively. The four areas under the curve calculations showed that both instruments possess a moderate and very similar degree of responsiveness in CLBP patients.

3.4. Concurrent validity

Only the COOP charts were administered twice during the 1 week baseline period ($n = 173$), and test-retest reliability was high (0.82) as assessed by the intraclass correlation coefficient. The coefficients for the individual nine dimensions varied between 0.51 and 0.79. The internal consistency coefficients (Chronbach's Alpha) were 0.77 and 0.8 for the COOP Charts and the SF-36, respectively. These test-retest and internal consistency coefficients are similar in magnitude to those found in other studies of both the COOP charts and the SF-36 (Beaton et al., 1997; Kinnersley et al., 1994; Nelson et al., 1990b). A high level of concurrent validity was present when correlating SF-36 and COOP Chart global scores at baseline ($r = 0.82$, 95% CI: 0.77, 0.86), and after 11 weeks of therapy ($r = 0.85$, 95% CI: 0.79, 0.89). When correlating the individual sets of dimensions from both instruments, the overall canonical correlation coefficients were 0.90 at baseline, and 0.91 at week 11, both statistically significant ($P < 0.0001$). The individual dimensions in the two instruments do not measure identical constructs, but six of them that describe very similar phenomena do show moderate to high correlations (0.52–0.86), all statistically significant at or below the 0.001 level (see Table 3).

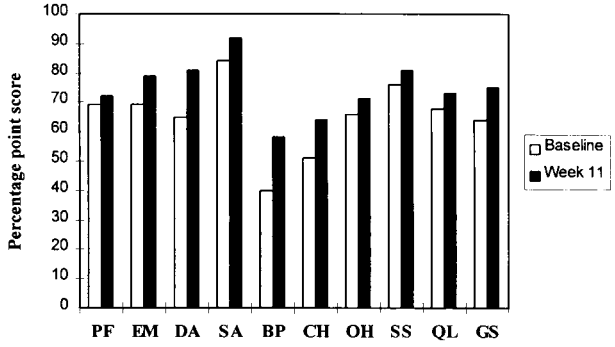

Fig. 2. COOP Chart dimension score profile in percentage points at baseline and at week 11. (Higher scores are better regardless of dimension). PF, Physical fitness; EM, Emotional; DA, Daily activity; SA, Social activity; BP, Bodily pain; CH, Change in health; OH, Overall health; SS, Social support; QL, Quality of Life; GS, Global score.

Table 2
Comparative responsiveness of the COOP charts and the SF-36 based on low back pain and disability as external criteria (EC)[a]

Outcome measure	External criterion: patient-rated low-back pain					
	Improved patients (≥ 10% points on EC)	Non-improved patients (− 5–5%points on EC)				
	Mean change SD (% points)	Mean change SD (% points)	Difference in mean change (% points)	Unpaired *t*-value	*P*-value	Effect size of difference in mean change (95% CI)
Low-back pain (EC)	32.63 ($n = 99$) 14.9					
COOP charts	13.48 ($n = 94$) 11.8	3.83 ($n = 20$) 8.6	9.65	3.46	0.0008	0.82 (0.32, 1.31)
SF-36	13.24 ($n = 98$) 12.5	3.77 ($n = 20$) 14.2	9.47	3.01	0.003	0.71 (0.22, 1.20)
	External criterion: patient-rated low-back disability					
	Improved patients (≥ 10% points on EC)	Non-improved patients (− 5–5% points on EC)				
	Mean change SD (% points)	Mean change SD (% points)	Difference in mean change (% points)	Mann–Whitney *Z* value	*P*-value	Effect size of difference in mean change (95% CI)
Low-back disability (EC)	24.92 ($n = 87$) 12.6					
COOP charts	13.80 ($n = 84$) 12.7	7.73 ($n = 19$) 6.5	6.07	2.32	0.02[b]	0.51 (0.01, 1.01)
SF-36	14.40 ($n = 87$) 12.8	6.76 ($n = 20$) 8.2	7.64	3.10	0.002[b]	0.61 (0.11, 1.10)

[a] The low-back pain, low-back disability and the health status instrument percentage point (%points) changes represent mean improvement in global scores after 11 weeks of therapy compared to baseline.
[b] Since group variances were significantly different, the non-parametric Mann–Whitney test was performed.

4. Discussion

There exists some uncertainty regarding optimal methods for assessing responsiveness of outcome instruments and for comparing competing instruments. Also, different methods of analysis yield different magnitudes of responsiveness (Wright and Young, 1997). A common method of assessing responsiveness is to compare scores of one or more new instruments before and after an efficacious treatment (Liang et al., 1990; Deyo et al., 1991). However, no clearly efficacious or 'gold standard' treatment currently exists for CLBP. Another method of quantifying responsiveness is to compare effect sizes defined as the change in mean pre/post treatment scores divided by the standard deviation of the change score (Cohen, 1988b; Liang et al., 1990). Some investigators advocate calculating the effect size using the standard deviation of the change score in stable patients to control for the non-specific changes in non-improved patients (Guyatt et al., 1987; Tuley et al., 1991). The standard deviation of the pre-treatment score has also been used for this purpose (Kazis et al., 1989).

We chose unpaired *t*-statistics (Guyatt et al., 1989) and the receiver operating characteristic (ROC) curves (Hanley and McNeil, 1982; Deyo et al., 1991) for assessment of instrument responsiveness. These methods allowed the determination of clinically important and statistically significant differences between improved and non-improved patients. Additionally, the ROC method enabled us to test whether the difference between the two instruments' responsiveness was of statistical significance (Hanley and McNeil, 1983).

There is currently no consensus on what constitutes a 'gold standard' in choice of EC to represent improvement when evaluating responsiveness of functional status instruments, and further research is needed in this area (Deyo and Centor, 1986; de Bruin et al., 1997). Examples of EC for establishing patient improvement in other responsiveness studies of specific clinical conditions are satisfaction with care (Stucki et al., 1995), change in health, (Beaton et al., 1997) pain improvement, clinician-rated improvement, return to normal activities (Deyo and Centor, 1986), and global perceived effect (Beurskens et al., 1996). Since there is no 'gold standard', a single generic outcome measure is unlikely to serve as an optimal EC (Deyo and Centor, 1986). Thus, we decided to use two complementary, reliable, valid and responsive low-back specific instruments as EC of improvement. The results were relatively consistent using both EC, and increased our confidence in the correctness of the calculated responsiveness.

Although the use of global or unweighted aggregate scores have been recommended for some health status instruments, it has not been advocated by the instrument developers of the two functional health status instruments in this study. We find, as have others (Beaton et al., 1997), that for the purpose of clinical trials, where separate analysis of multiple outcomes may seriously reduce statistical

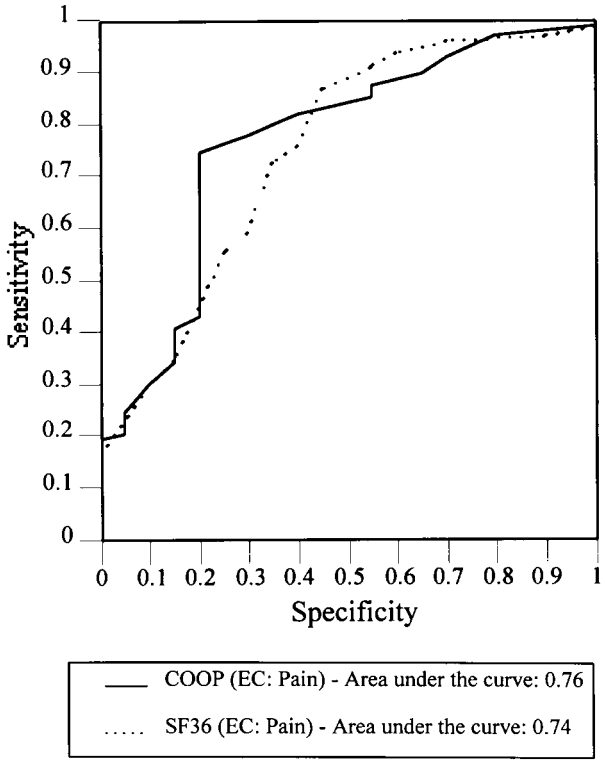

Fig. 3. Receiver operating characteristic (ROC) curves for COOP and SF-36 global change scores (baseline to week 11) based on patient-rated low-back pain as external criterion (EC), with ≥ 10 percentage points representing improved and − 5–5 percentage points non-improved patients.

power, the global score (a pooled index) can be advantageous. The global score may also be clinically meaningful if it can be established, as in this study, that responsiveness to change in the studied clinical condition is adequate. However, the magnitude of change representing clinical importance remains an open question. John Ware, the original developer of the SF-36, has suggested that a change in individual domain scores of 6–8 percentage points is clinically important (pers. commun.). Certainly, the 9–11 percentage points improvement in global scores in the LBP patients in this study translates into an important clinical change and corroborates well with the 20–25 percentage points improvement in the condition-specific patient-rated low-back pain and disability. It should be noted that a change of approximately 6 percentage points in global score, the best cut-off point for both instruments in this study, was capable of distinguishing between improved and non-improved CLBP patients.

Similar global score changes may represent very different individual dimension profile changes for different clinical conditions. The SF-36 was used as one of the primary outcome measures in a RCT on spinal manipulation versus Amitriptyline for chronic muscle tension headache (Boline et al., 1995). Corresponding with the difference in headache pain ratings, the group difference of 5 percentage points in the SF-36 global scores was primarily attributed to three of the nine dimensions: health transition, general health perception and emotional problems in relation to role-function. In contrast, in the present study on LBP patients, the SF-36's bodily pain and physical problems in relation to role-function, and the similar COOP Chart dimensions, bodily pain and daily activity were the dimensions to show the largest change. This appears to make intuitive sense in terms of the clinical characteristics of these two different disorders.

The results of our study showed sufficient and very similar responsiveness of both instruments. Consistent with our results, Jenkinson et al. (1995), using effect size statistics, also found very similar responsiveness of the COOP charts and SF-36 in 129 patients participating in a RCT on two different surgical procedures for inguinal hernias. It must be noted that our study sample was treated in a primary care setting and consisted of chronic moderately severe back pain patients of whom a relatively small percentage were not working at full capacity. Therefore, additional evaluation of these instruments is needed before it is known whether the results of this study apply to more severely afflicted back pain patients, who would often be receiving care in tertiary chronic back pain centers.

Our study, as have others (Landgraf, 1990; Nelson et al., 1990b; McHorney et al., 1992; Anderson et al., 1993), showed that the individual dimensions or domains of the SF-36 and the COOP charts are not identical, but do create

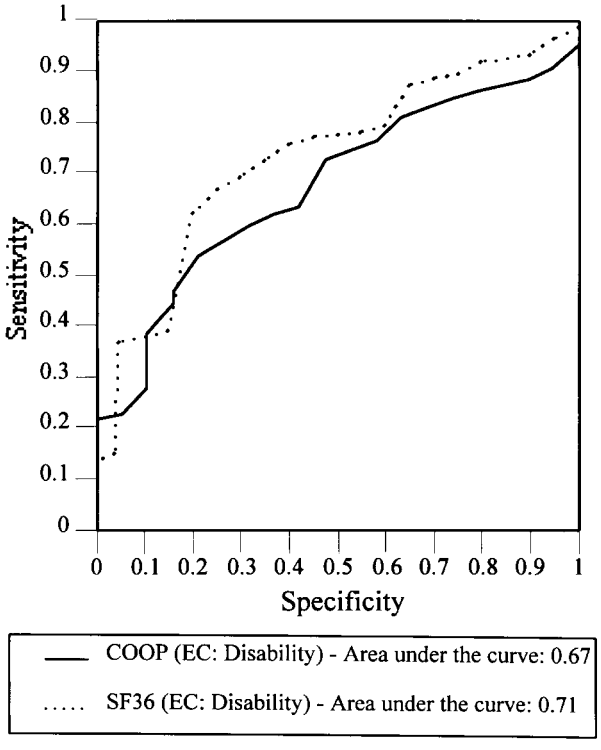

Fig. 4. Receiver operating characteristic (ROC) curves for COOP and SF-36 global change scores (baseline to week 11) based on patient-rated disability as external criterion (EC), with ≥ 10 percentage points representing improved and − 5–5 percentage points non-improved patients.

Table 3
Correlation of the six similar dimensions (bold) of the COOP and SF-36 instruments at baseline ($n = 173$)[a]

	COOP-PF	COOP-DA	COOP-SA	COOP-BP	COOP-E	COOP-OH
SF-36-PF	**0.56**	0.49	0.34	0.26	0.13	0.36
SF-36-RP	0.26	**0.52**	0.42	0.42	0.25	0.35
SF-36-SF	0.22	0.58	**0.86**	0.22	0.49	0.54
SF-36-BP	0.18	0.54	0.35	**0.61**	0.29	0.28
SF-36-MH	0.11	0.38	0.54	0.15	**0.74**	0.44
SF-36-GH	0.32	0.40	0.42	0.20	0.24	**0.72**

[a] COOP dimensions: PF, Physical fitness; DA, Daily activity; SA, Social activity; BP, Bodily pain; E, Emotional; OH, Overall health. SF-36 dimensions: PF, Physical functioning; RP, Role function physical; SF, Social functioning; BP, Bodily pain; MH, Mental health; GH, General health.

somewhat different patient profiles. However, six of the total of nine domains describing very similar constructs are moderately to highly correlated. The construct of functional health status as reflected in the global scores and the canonical correlation analysis are also highly correlated.

Although not systematically assessed, our estimates of mean time used by patients to complete the instruments correspond with that reported in the literature, 2–5 min for the COOP charts (Nelson et al., 1987) and 8–15 min for the SF-36 (Weinberger et al., 1991; Beaton et al., 1997). The COOP charts are faster and easier to score (Jenkinson et al., 1995), but illustrations accompanying the text in the COOP charts prevent them from being used as intended when administered over the phone. However, at least two studies involving different types of patients have shown that administering the COOP charts without the pictograms do not seem to affect outcomes (Larson et al., 1992; Kempen et al., 1997). This is not an issue with the SF-36.

Among the short functional health status instruments, the SF-36 and the COOP charts seem to have the most widespread use internationally (Landgraf and Nelson, 1992; Wasson et al., 1992; Anderson et al., 1993; Bruusgaard et al., 1993; Lam et al., 1994), and cross-cultural validation studies are currently underway for both instruments (Anderson et al., 1993). The SF-36 is being assessed in 15 countries for the purpose of validation and establishing norms to be used in future international clinical trials (Anderson et al., 1993). In 1988, the World Organization of General Practitioners/Family Physicians (WONCA) chose the COOP charts as the basis for developing an international system for assessing functional health status (Landgraf and Nelson, 1992). As a result of these studies, revised and shorter versions of both instruments are being developed and tested (Ware et al., 1996). Although a substantial amount of normative and condition-specific profiles now exists especially for the SF-36 (Garratt et al., 1993; Jenkinson et al., 1993), both instruments' responsiveness in various specific patient populations requires further research.

5. Conclusions

Using two complementary methods of analysis, the COOP charts and the SF-36 show moderate and very similar responsiveness in CLBP patients. The construct of functional health status, as reflected in the global scores of the two instruments, are highly correlated. Six of the total nine domains of each of the two instruments are moderately to highly correlated. Both instruments seem equally suitable for use as primary outcome measures in clinical trials on CLBP. The COOP charts are faster to fill out and score.

Acknowledgements

We wish to thank R. Evans and M. Haas for the constructive reviews of this manuscript.

References

Anderson RT, Aaronson NK, Wilkin D. Critical review of the international assessments of health-related quality of life. Qual Life Res 1993;2:369–395.

Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79–93.

Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71–76.

Boline PD, Kassak K, Bronfort G, Nelson C, Anderson AV. Spinal manipulation vs. amitriptyline for the treatment of chronic tension-type headaches: a randomized clinical trial. J Manipulative Physiol Ther 1995;18:148–154.

Bronfort G, Goldsmith CH, Nelson CF, Boline PD, Anderson AV. Trunk exercise combined with spinal manipulative or NSAID therapy for chronic low back pain: a randomized, observer-blinded clinical trial. J Manipulative Physiol Ther 1996;19:570–582.

Bruusgaard D, Nessioy I, Rutle O, Furuseth K, Natvig B. Measuring functional status in a population survey. The Dartmouth COOP functional health assessment charts/WONCA used in an epidemiological study. Fam. Pract. 1993;10:212–218.

Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988a. p. 381.

Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988b. pp. 8–14.

de Bruin AF, Diederiks JPM, de Witte LP, Stevens FCJ, Philipsen H. Assessing the responsiveness of a functional status measure: the sickness impact profile versus the SIP68. J Clin Epidemiol 1997;50:529–540.

Deyo RA. Comparative validity of the sickness impact profile and shorter scales for functional assessment in low-back pain. Spine 1986;11:951–954.

Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39:897–906.

Deyo RA, Diehl AK. Measuring physical and psychosocial function in patients with low-back pain. Spine 1983;8:635–642.

Deyo RA, Walsh NE, Martin DC, Schoenfeld LS, Ramamurthy S. A controlled trial of transcutaneous electrical nerve stimulation (TENS) and exercise for chronic low back pain. N Engl J Med 1990;322:1627–1634.

Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12:142S–158S.

Deyo RA, Andersson G, Bombardier C, Cherkin DC, Keller RB, Lee CK, Liang MH, Lipscomb B, Shekelle P, Spratt KF, et al. Outcome measures for studying patients with low back pain. Spine 1994;19:2032S–2036S.

Garratt AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT. The SF36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? Br Med J 1993;306:1440–1444.

Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee J Rheumatol 1993;20:561–565.

Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171–178.

Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol 1989;42:403–408.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the saxme cases. Radiology 1983;148:839–843.

Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando, FL: Academic Press, 1985. pp. 2–46.

Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–415.

Jaeschke R, Singer J, Guyatt GH. A comparison of seven-point and visual analogue scales. Data from a randomized trial. Control Clin Trials 1990;11:43–51.

Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age. Br Med J 1993;306:1437–1440.

Jenkinson C, Lawrence K, McWhinnie D, Gordon J. Sensitivity to change of health status measures in a randomized controlled trial: comparison of the COOP charts and the SF-36. Qual life Res 1995;4:47–52.

Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. J Clin Epidemiol 1994;47:81–87.

Kachigan SK. An interdisciplinary introduction to univariate and multivariate methods, anonymous statistical analysis. New York: Radius Press, 1986. pp. 234-237.

Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178–S189.

Kempen GI, van Sonderen E, Sanderman R. Measuring health status with the Dartmouth COOP charts in low-functioning elderly. Do the illustrations affect the outcomes? Qual Life Res 1997;6:323–328.

Kinnersley P, Peters T, Stott N. Measuring functional health status in primary care using the COOP-WONCA charts: acceptability, range of scores, construct validity, reliability and sensitivity to change. Br J Gen Pract 1994;44:545–549.

Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GM, Hofhuizen DM, Houben JP, Knipschild PG. The effectiveness of manual therapy, physiotherapy, and treatment by the general practitioner for nonspecific back and neck complaints. A randomized clinical trial. Spine 1992;17:28–35.

Lam CL, van Weel C, Lauder IJ. Can the Dartmouth COOP/WONCA charts be used to assess the functional status of Chinese patients? Fam Pract 1994;11:85–94.

Landgraf JM. Assession function: does it really make a difference? A preliminary evaluation of the acceptability and utility of the COOP function charts. Anonymous functional status measurement in primary care. In: Lipkin M, editor. New York: Springer, 1990.

Landgraf JM, Nelson EC. Summary of the WONCA/COOP international health assessment field trial. The Dartmouth COOP primary care network. Aust Fam Physician 1992;21:255–257.

Larson CO, Hays RD, Nelson EC. Do the pictures influence scores on the Dartmouth COOP Charts? Qual Life Res 1992;1:247–249.

Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632–642.

Little RJA, Rubin D. Statistical analysis with missing data. New York: Wiley, 1987. pp. 21–37.

McHorney CA, Ware Jr. JE, Rogers W, Raczek AE, Lu JF. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. Results from the Medical Outcomes Study. Med Care 1992;30:MS253–MS265.

Meyboom-de Jong B, Smith RJA. Studies with the Dartmouth COOP charts in general practice: comparison with the Nottingham Health Profile and the General Health Questionnaire. In: Lipkin M, editor. Functional status measurement in primary care, New York: Springer, 1990. pp. 132–149.

Nelson E, Conger B, Douglass R, Gephart D, Kirk J, Page R, Clark A, Johnson K, Stone K, Wasson J, Zubkoff M. Functional health status levels of primary care patients. J Am Med Assoc 1983;249:3331–3338.

Nelson E, Wasson J, Kirk J, Keller A, Clark D, Dietrich A, Stewart A, Zubkoff M. Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. J Chronic Dis 1987;40(Suppl 1):55S–69S.

Nelson E, Wasson J, Kirk J, et al. Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. The Medical Outcomes Study Instrument (MOSI)–use of a new health status measure in Britain. J Chronic Dis Fam Pract 1990a;7:205–218.

Nelson EC, Landgraf JM, Hays RD, Kirk JW, Wasson AK, Zubkoff M. The COOP function charts: a system to measure patient function in physician's offices. Functional status measurement in primary care. In: Lipkin M, editor. New York: Springer, 1990b. pp. 97–131.

Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8:141–144.

Shigemoto H. A trial of the Dartmouth COOP charts in Japan. Functional status measurement in primary care. In: Lipkin M, editor. New York: Springer, 1990. pp. 181–190.

Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, McGlynn EA, Ware Jr. JE. Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study. J Am Med Assoc 1989;262:907–913.

Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369–1378.

Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. J Clin Epidemiol 1991;44:417–421.

van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the rehabilitation activities profile and the Barthel Index. J Clin Epidemiol 1996;49:39–44.

Ware Jr J, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220–233.

Wasson J, Keller A, Rubenstein L, Hays R, Nelson E, Johnson D. Benefits

and obstacles of health status assessment in ambulatory settings. The clinician's point of view. The Dartmouth Primary Care COOP Project. Med Care 1992;30:MS42–MS49.
Weinberger M, Samsa GP, Hanlon JT, Schmader K, Doyle ME, Cowper PA, Uttech KM, Cohen HJ, Feussner JR. An evaluation of a brief health status measure in elderly veterans. J Am Geriatr Soc 1991;39:691–694.
Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239–246.